Revisiting inference after prediction

Keshav Motwani   KMOTWANI@UW.EDU
Department of Biostatistics
University of Washington
Seattle, WA

Daniela Witten   DWITTEN@UW.EDU
Departments of Biostatistics and Statistics
University of Washington
Seattle, WA

Editor: Mladen Kolar

Abstract

Recent work has focused on the very common practice of prediction-based inference: that is, (i) using a pre-trained machine learning model to predict an unobserved response variable, and then (ii) conducting inference on the association between that predicted response and some covariates. As pointed out by Wang et al. (2020), applying a standard inferential approach in (ii) does not accurately quantify the association between the unobserved (as opposed to the predicted) response and the covariates. In recent work, Wang et al. (2020) and Angelopoulos et al. (2023) propose corrections to step (ii) in order to enable valid inference on the association between the unobserved response and the covariates. Here, we show that the method proposed by Angelopoulos et al. (2023) successfully controls the type 1 error rate and provides confidence intervals with correct nominal coverage, regardless of the quality of the pre-trained machine learning model used to predict the unobserved response. However, the method proposed by Wang et al. (2020) provides valid inference only under very strong conditions that rarely hold in practice: for instance, if the machine learning model perfectly estimates the true regression function in the study population of interest.

1. Introduction

Rapid recent progress in the field of machine learning has enabled the development of complex and high-quality machine learning models to predict a response variable of interest. This is particularly attractive in settings where future measurement of this response variable is prohibitively expensive or impossible. For example, instead of performing expensive experiments to determine a protein's structure, it is possible to obtain high-quality structural predictions using AlphaFold (Jumper et al., 2021). Similarly, in developing countries, determining the true cause of death may be impossible; instead, one might predict the cause of death on the basis of a "verbal autopsy" (Clark et al., 2015; Khoury et al., 1999). In the context of gene expression data, it is infeasible to conduct experiments in every possible tissue type, and so instead a machine learning model can be applied to predict gene expression in a tissue type of interest (Ellis et al., 2018; Gamazon et al., 2015; Gusev et al., 2019).

In this paper, we will use the notation $\hat{f} : \mathcal{Z} \to \mathcal{Y}$ to denote a pre-trained machine learning model that maps from $\mathcal{Z}$, the space of the predictors Z, to $\mathcal{Y}$, the space of the response variable Y. We assume that the operating characteristics of $\hat{f}(\cdot)$ in the study population of interest are unknown to the end-user, and the data used to fit $\hat{f}(\cdot)$ are unavailable. Thus, in what follows, we will treat $\hat{f}(\cdot)$ as a "black box" function.

In important recent papers, Wang et al. (2020) and Angelopoulos et al. (2023) consider the practice of *prediction-based inference*[1]: that is, of quantifying the association between the response Y and some covariates X using realizations not of (Y, X), but rather, of $(\hat{f}(Z), X)$. Such an approach is attractive in cases where the association between Y and X is of interest, and both $\hat{f}(\cdot)$ and a large sample of realizations of predictors Z and covariates X are available, but realizations of Y are expensive or otherwise unavailable. Throughout, we will use Z to denote the *predictors* in the machine learning model $\hat{f}(\cdot)$, and X to denote the *covariates* whose association with Y is of interest. In many settings, the covariate variables X may be a subset of the predictor variables Z, or may be identical to Z, but this is not necessarily the case.

Wang et al. (2020) point out that a naive approach to prediction-based inference that simplistically interprets the association between $(\hat{f}(Z), X)$ as the association between (Y, X) is problematic from a statistical perspective. For instance, regressing $\hat{f}(Z)$ onto X using least squares does not lead to valid inference on the association between Y and X: e.g. it leads to hypothesis tests that fail to control type 1 error, and confidence intervals that do not achieve the nominal coverage. See Box 1.

Box 1: *Naive approach to prediction-based inference.* We are given a (pre-trained) prediction function $\hat{f}(\cdot) : \mathcal{Z} \to \mathcal{Y}$, and an unlabeled dataset $\mathbf{z}_{\mathrm{unlab}}$ representing realizations from Z. As pointed out by Wang et al. (2020), the naive approach displayed here does not allow for valid inference on the association between Y and X.

Step 1: Compute $\hat{\mathbf{y}}_{\mathrm{unlab}} = \hat{f}(\mathbf{z}_{\mathrm{unlab}})$.

Step 2: Conduct inference on the association between Y and X, using $\hat{\mathbf{y}}_{\mathrm{unlab}}$ and $\mathbf{x}_{\mathrm{unlab}}$ as data, without accounting for the fact that $(\hat{\mathbf{y}}_{\mathrm{unlab}}, \mathbf{x}_{\mathrm{unlab}})$ is not a sample from the distribution of (Y, X).

Wang et al. (2020) and Angelopoulos et al. (2023) propose creative solutions to overcome this issue. They assume that in addition to a large unlabeled dataset $(\mathbf{z}_{\mathrm{unlab}}, \mathbf{x}_{\mathrm{unlab}})$, they also have access to a (relatively small) labeled dataset $(\mathbf{y}_{\mathrm{lab}}, \mathbf{z}_{\mathrm{lab}}, \mathbf{x}_{\mathrm{lab}})$. We focus on the case where the labeled and unlabeled data are independent and identically distributed samples from the same study population of interest, though Angelopoulos et al. (2023) also extend beyond this setting. See Box 2. We emphasize that the goal is to quantify association between Y and X.

---

1. Wang et al. (2020) and Angelopoulos et al. (2023) refer to this practice as *post-prediction inference* and *prediction-powered inference*, respectively; to unify terminology, here we refer to it as *prediction-based inference*.

> Box 2: *Setting of prediction-based inference.*
>
> - *Given:* A pre-trained machine learning model $\hat{f} : \mathcal{Z} \to \mathcal{Y}$.
>
> - *Goal:* To quantify the association between a response $Y \in \mathcal{Y}$ and covariates $X \in \mathcal{X}$.
>
> - *Data:* A (relatively small) labeled dataset $(\mathbf{y}_{\text{lab}}, \mathbf{z}_{\text{lab}}, \mathbf{x}_{\text{lab}})$ consisting of i.i.d. realizations of (Y, Z, X), and a (large) unlabeled dataset $(\mathbf{z}_{\text{unlab}}, \mathbf{x}_{\text{unlab}})$ consisting of i.i.d. realizations of (Z, X). Both are drawn from the same study population.

One simple (and valid) option is to quantify association between Y and X using only the labeled data $(\mathbf{y}_{\text{lab}}, \mathbf{x}_{\text{lab}})$. However, this approach entirely discards the vast amount of unlabeled data $(\mathbf{z}_{\text{unlab}}, \mathbf{x}_{\text{unlab}})$. Intuitively, if the prediction function $\hat{f}(\cdot)$ is nearly perfect on our population of interest, then using $(\hat{f}(\mathbf{z}_{\text{unlab}}), \mathbf{x}_{\text{unlab}})$ in addition to $(\mathbf{y}_{\text{lab}}, \mathbf{x}_{\text{lab}})$ will aid our efforts to quantify the association between Y and X. By contrast, if $\hat{f}(Z)$ is a very poor prediction of Y on our population of interest, then using $(\hat{f}(\mathbf{z}_{\text{unlab}}), \mathbf{x}_{\text{unlab}})$ in addition to $(\mathbf{y}_{\text{lab}}, \mathbf{x}_{\text{lab}})$ may hinder our efforts. Our goal is valid quantification of the association between Y and X, *regardless of the quality of $\hat{f}(\cdot)$ on the population of interest.*

To achieve this goal, Wang et al. (2020) propose to (Step 1') model the association between Y and $\hat{f}(Z)$ using the labeled dataset $(\mathbf{y}_{\text{lab}}, \hat{f}(\mathbf{z}_{\text{lab}}))$, and then (Step 2') incorporate the model in Step 1' to conduct inference between Y and X using the unlabeled data $(\hat{f}(\mathbf{z}_{\text{unlab}}), \mathbf{x}_{\text{unlab}})$. See Box 3.

> Box 3: *Wang et al. (2020)'s proposal to correct prediction-based inference.*
>
> Step 1': Model the association between Y and $\hat{f}(Z)$ using the labeled data $(\mathbf{y}_{\text{lab}}, \hat{f}(\mathbf{z}_{\text{lab}}))$.
>
> Step 2': Incorporate the model from Step 2' to conduct inference on the association between Y and X using the unlabeled data $(\hat{f}(\mathbf{z}_{\text{unlab}}), \mathbf{x}_{\text{unlab}})$. To do this, bootstrap and analytical approaches are proposed.

By contrast, Angelopoulos et al. (2023) propose to de-bias the estimates obtained using the unlabeled data using information from the labeled data. In the case of estimands that are linear in Y, they (Step 1") compute the difference between the estimate of the parameter of interest obtained using $(\hat{f}(\mathbf{z}_{\text{lab}}), \mathbf{x}_{\text{lab}})$ and the estimate obtained using $(\mathbf{y}_{\text{lab}}, \mathbf{x}_{\text{lab}})$. They then (Step 2") correct the estimate obtained using the unlabeled dataset $(\hat{f}(\mathbf{z}_{\text{unlab}}), \mathbf{x}_{\text{unlab}})$ by this amount. See Box 4. Angelopoulos et al. (2023) also propose a more general framework for estimands that minimize the expectation of a general loss function, though we focus on the linear case in this paper for simplicity.

In this paper, we investigate these two proposals. In Section 2, we ask a fundamental question: what parameter is each proposal targeting? We see that the proposal of Angelopoulos et al. (2023) targets the parameter of interest, whereas that of Wang et al. (2020) does not. In Sections 3 and 4, we investigate the empirical consequences of our

> Box 4: *Angelopoulos et al. (2023)'s proposal to correct prediction-based inference (in the special case of estimands that are linear in Y).*
>
> Step 1": Compute the difference between the estimate of the parameter of interest obtained using $(\hat{f}(\mathbf{z}_{\text{lab}}), \mathbf{x}_{\text{lab}})$ and the estimate obtained using $(\mathbf{y}_{\text{lab}}, \mathbf{x}_{\text{lab}})$.
>
> Step 2": Correct the parameter estimate obtained using the unlabeled dataset $(\hat{f}(\mathbf{z}_{\text{unlab}}), \mathbf{x}_{\text{unlab}})$ by the difference computed in Step 1".

findings from Section 2. These empirical investigations paint a clear picture: namely, that failure to target the correct parameter has substantial statistical consequences for the proposal of Wang et al. (2020), in the form of hypothesis tests that fail to control the Type 1 error, and confidence intervals that fail to attain the nominal coverage. The proposal of Angelopoulos et al. (2023) does not suffer these consequences, as it targets the correct parameter. We close with a discussion in Section 5.

In this paper, we use capitals to represent a random variable and lower case to represent its realization. Vectors of length equal to the number of observations, or matrices whose rows correspond to the observations, are in bold.

2. What parameter is each method targeting?

For concreteness, suppose that we would have fit a linear regression model on realizations of (Y, X) using least squares, had a large number of realizations of (Y, X) been available. Therefore, our goal is to conduct inference on the population parameter

$$\beta^* = \arg\min_{\beta} \text{E}[(Y - \beta^\top X)^2] = \text{E}[XX^\top]^{-1}\,\text{E}[XY]. \tag{1}$$

The naive method (Box 1) and the proposals of Wang et al. (2020) and Angelopoulos et al. (2023) use a test statistic of the form

$$T_j = \frac{\hat{\beta}_j - \beta_j^*}{\widehat{\text{SE}}(\hat{\beta}_j)}$$

for testing or constructing confidence intervals for β_j^*. They rely on it having a known distribution, or converging in distribution to a known distribution with increasing sample size.

However, if $\hat{\beta}_j \overset{p}{\nrightarrow} \beta_j^*$ and $\widehat{\text{SE}}(\hat{\beta}_j)$ goes to 0 as n_{lab} and n_{unlab} increase, then T_j does not converge in distribution (see Appendix A for a formal statement of this). Therefore, consistency of $\hat{\beta}_j$ for β_j^* is a necessary condition for valid inference using this approach. We now investigate whether this is the case.

2.1 The general case for an arbitrary prediction model $\hat{f}$

We first consider the naive approach, as defined in Box 1. Fitting a linear model with least squares would result in $\hat{\beta}_{\text{naive}} = (\mathbf{x}_{\text{unlab}}^\top \mathbf{x}_{\text{unlab}})^{-1}\mathbf{x}_{\text{unlab}}^\top \hat{f}(\mathbf{z}_{\text{unlab}})$. We can see that $\hat{\beta}_{\text{naive}} \overset{p}{\rightarrow}$

Algorithm 1 Bootstrap correction of Wang et al. (2020). The goal is to conduct inference on the association between Y and X.

1. Use $(\mathbf{y}_{\mathrm{lab}}, \hat{f}(\mathbf{z}_{\mathrm{lab}}))$ to fit the "relationship model" $Y|\hat{f}(Z) \sim K(\hat{f}(Z), \phi)$, yielding $\hat{\phi}$.

2. For $b = 1, \ldots, B$:

    2.1. Sample unlabeled observations with replacement to obtain $\mathbf{z}_{\mathrm{unlab}}^b$ and $\mathbf{x}_{\mathrm{unlab}}^b$.

    2.2. Sample outcomes $\tilde{\mathbf{y}}^b|\hat{f}(\mathbf{z}_{\mathrm{unlab}}^b)$ from the relationship model $K(\hat{f}(\mathbf{z}_{\mathrm{unlab}}^b), \hat{\phi})$.

    2.3. Use $(\tilde{\mathbf{y}}^b, \mathbf{x}_{\mathrm{unlab}}^b)$ to fit a "regression model" for the relationship between Y and X, and record the coefficient estimate $\hat{\beta}^b$ and model-based standard error $\hat{s}^b$.

3. Compute the point estimate $\hat{\beta} = \mathrm{median}\{\hat{\beta}^1, \ldots, \hat{\beta}^B\}$.

4. Compute the "nonparametric" standard error $\widehat{\mathrm{SE}}(\hat{\beta}) = \mathrm{SD}\{\hat{\beta}^1, \ldots, \hat{\beta}^B\}$.

5. Compute the "parametric" standard error $\widehat{\mathrm{SE}}(\hat{\beta}) = \mathrm{median}\{\hat{s}^1, \ldots, \hat{s}^B\}$.

$\mathrm{E}[XX^\top]^{-1}\,\mathrm{E}[X\hat{f}(Z)]$ as n_{unlab} increases, which is *not* equal to $\beta^* = \mathrm{E}[XX^\top]^{-1}\,\mathrm{E}[XY]$ in general. In fact, viewing $\hat{f}(\cdot)$ as a black-box function with unknown operating characteristics in the study population, we see that the quantity $\mathrm{E}[XX^\top]^{-1}\,\mathrm{E}[X\hat{f}(Z)]$ does not even involve the response, Y; therefore, the parameter targeted by the naive method is not of any scientific interest.

Now, we consider the bootstrap variant of the proposal of Wang et al. (2020), which is introduced in Box 3. A detailed description of these proposals are presented in Algorithm 1. We take Step 2.3 to involve a least squares regression. Note that Step 2.2 of Algorithm 1 involves sampling observations $\tilde{\mathbf{y}}^b$ from $K(\hat{f}(Z), \hat{\phi})$ for use in fitting a regression model in Step 2.3, giving us an estimate $\hat{\beta}_{\mathrm{Wang}}^{\mathrm{bootstrap},b} = (\mathbf{x}_{\mathrm{unlab}}^\top \mathbf{x}_{\mathrm{unlab}})^{-1} \mathbf{x}_{\mathrm{unlab}}^\top \tilde{\mathbf{y}}^b$. Thus $\hat{\beta}_{\mathrm{Wang}}^{\mathrm{bootstrap},b} \xrightarrow{p} \mathrm{E}[XX^\top]^{-1}\,\mathrm{E}[XK(\hat{f}(Z), \hat{\phi})]$ as n_{unlab} increases, where we slightly abuse notation by letting $K(\hat{f}(Z), \hat{\phi})$ denote a random variable with distribution $K(\hat{f}(Z), \hat{\phi})$. In general, $\mathrm{E}[XX^\top]^{-1}\,\mathrm{E}[XK(\hat{f}(Z), \hat{\phi})]$ is not equal to the parameter of interest $\beta^* = \mathrm{E}[XX^\top]^{-1}\,\mathrm{E}[XY]$. Note that both the "parametric" and "non-parametric" bootstrap corrections suffer from this issue.

The analytic variant of the proposal of Wang et al. (2020) adjusts the naive estimate by the coefficient of $\hat{f}(Z)$ in a regression of Y onto $\hat{f}(Z)$, with an intercept, using the labeled data[2]; that is, the estimate takes the form

$$\hat{\beta}_{\mathrm{Wang}}^{\mathrm{analytical}} = \frac{\widehat{\mathrm{Cov}}(\mathbf{y}_{\mathrm{lab}}, \hat{f}(\mathbf{z}_{\mathrm{lab}}))}{\widehat{\mathrm{Var}}(\hat{f}(\mathbf{z}_{\mathrm{lab}}))} \hat{\beta}_{\mathrm{naive}}. \tag{2}$$

Thus as n_{lab} and n_{unlab} increase,

$$\hat{\beta}_{\mathrm{Wang}}^{\mathrm{analytical}} \xrightarrow{p} \frac{\mathrm{Cov}(Y, \hat{f}(Z))}{\mathrm{Var}(\hat{f}(Z))} \mathrm{E}[XX^\top]^{-1}\,\mathrm{E}[X\hat{f}(Z)],$$

which is not again not equal to β^* in general.

---

2. The expression for $\hat{\beta}_{\mathrm{Wang}}^{\mathrm{analytical}}$ given in (2) is implemented in Wang et al. (2020)'s code. Their publication involves a slightly different expression for $\hat{\beta}_{\mathrm{Wang}}^{\mathrm{analytical}}$, which also does not converge to the parameter of interest β^* for a similar reason – we show this in Appendix B.

This highlights a cause for concern about the proposals of Wang et al. (2020): namely, that the wrong parameter is being targeted. This calls into question whether the inference that they propose will achieve the desired statistical guarantees; we investigate this issue further in the next two sections.

Finally, we turn to the proposal of Angelopoulos et al. (2023), which is introduced in Box 4. In the case of linear regression, their estimate takes the form

$$\hat{\beta}_{\text{Angelopoulos}} = (\mathbf{x}_{\text{unlab}}^\top \mathbf{x}_{\text{unlab}})^{-1} \mathbf{x}_{\text{unlab}}^\top \hat{f}(\mathbf{z}_{\text{unlab}}) + \left\{ (\mathbf{x}_{\text{lab}}^\top \mathbf{x}_{\text{lab}})^{-1} \mathbf{x}_{\text{lab}}^\top \mathbf{y}_{\text{lab}} - (\mathbf{x}_{\text{lab}}^\top \mathbf{x}_{\text{lab}})^{-1} \mathbf{x}_{\text{lab}}^\top \hat{f}(\mathbf{z}_{\text{lab}}) \right\}.$$

As n_{lab} and n_{unlab} increase, we see that

$$\hat{\beta}_{\text{Angelopoulos}} \xrightarrow{p} \mathrm{E}[XX^\top]^{-1} \mathrm{E}[X\hat{f}(Z)] + \left\{ \mathrm{E}[XX^\top]^{-1} \mathrm{E}[XY] - \mathrm{E}[XX^\top]^{-1} \mathrm{E}[X\hat{f}(Z)] \right\} = \beta^*,$$

so the proposal of Angelopoulos et al. (2023) correctly targets the desired quantity.

2.2 An extreme setting where all methods target the correct quantity

We now consider an extreme setting where the prediction model *exactly* equals the true regression function: that is, $\hat{f}(z) = \mathrm{E}[Y|Z = z]$. We also assume that X is contained within Z: that is, $Z = (X, Z^{(2)})$ for some $Z^{(2)}$. This is a reasonable assumption in practice, since the covariate of interest is likely also a predictor in the machine learning model. In this extreme setting, the naive method targets

$$\begin{aligned} \mathrm{E}[XX^\top]^{-1} \mathrm{E}[X\hat{f}(Z)] &= \mathrm{E}[XX^\top]^{-1} \mathrm{E}\left[X\,\mathrm{E}[Y|Z]\right] \\ &= \mathrm{E}[XX^\top]^{-1} \mathrm{E}[XY] \\ &= \beta^*. \end{aligned}$$

In other words, it targets the correct parameter of interest.

We will now show that the bootstrap methods of Wang et al. (2020) can similarly target the correct parameter in this extreme setting, for example, if $K(\hat{f}(Z), \hat{\phi})$ is defined by fitting a generalized additive model (GAM) to $(Y, \hat{f}(Z))$ and adding mean-zero noise, as in Wang et al. (2020). That is,

$$K(\hat{f}(Z), \hat{\phi}) \overset{d}{=} \hat{g}(\hat{f}(Z)) + \epsilon, \tag{3}$$

where ϵ is mean-zero noise and $\hat{g}(\cdot)$ is the fitted GAM.

The fitted GAM $\hat{g}(\cdot)$ takes the form

$$\hat{g}(\cdot) \approx \arg\min_{g} \mathrm{E}\left[(Y - g(\hat{f}(Z)))^2 \right] \tag{4}$$

$$= \arg\min_{g} \mathrm{E}\left[(Y - g(\mathrm{E}[Y|Z]))^2 \right] \tag{5}$$

$$= \arg\min_{g} \mathrm{E}\left[\mathrm{E}\left[(Y - g(\mathrm{E}[Y|Z]))^2 |Z \right] \right]. \tag{6}$$

The approximation in (4) holds if the labeled sample size is sufficiently large. Equation 5 is a consequence of the extreme assumption that $\hat{f}(z) = \mathrm{E}[Y|Z = z]$. Equation 6 follows from iterated expectations.

It is not hard to see that (6) is minimized when $g(\mathrm{E}[Y|Z]) = \mathrm{E}[Y|Z]$, i.e., $\hat{g}(\cdot)$ is approximately the identity function. Combining this with the extreme assumption that $\hat{f}(z) = \mathrm{E}[Y|Z = z]$, (3) leads to

$$K(\hat{f}(Z), \hat{\phi}) \overset{d}{\approx} \mathrm{E}[Y|Z] + \epsilon. \tag{7}$$

Now, recall from Section 2.1 that the parameter targeted by Wang et al. (2020) is $\mathrm{E}[XX^\top]^{-1} \mathrm{E}[XK(\hat{f}(Z), \hat{\phi})]$. We observe that

$$\mathrm{E}[XX^\top]^{-1} \mathrm{E}[XK(\hat{f}(Z), \hat{\phi})] = \mathrm{E}[XX^\top]^{-1} \mathrm{E}[X \, \mathrm{E}[K(\hat{f}(Z), \hat{\phi})|Z]] \tag{8}$$

$$\approx \mathrm{E}[XX^\top]^{-1} \mathrm{E}[X \, \mathrm{E}[Y|Z]] \tag{9}$$

$$= \mathrm{E}[XX^\top]^{-1} \mathrm{E}[XY] \tag{10}$$

$$= \beta^*.$$

Here, (8) and (10) follow from iterated expectations since $Z = (X, Z^{(2)})$, and (9) follows from (7). With a large enough labeled dataset, this approximation will hold almost exactly.

The analytical method of Wang et al. also targets the correct parameter in this setting, since the naive method targets the correct parameter and

$$\frac{\mathrm{Cov}(Y, \hat{f}(Z))}{\mathrm{Var}(\hat{f}(Z))} = \frac{\mathrm{Cov}(Y, \mathrm{E}[Y|Z])}{\mathrm{Var}(\mathrm{E}[Y|Z])} = \frac{\mathrm{Var}(\mathrm{E}[Y|Z])}{\mathrm{Var}(\mathrm{E}[Y|Z])} = 1.$$

Therefore, we have seen that under a very extreme assumption that $\hat{f}(z) = \mathrm{E}[Y|Z = z]$, the methods of Wang et al. (2020) will target (nearly) the correct parameter. However, in general, this assumption is not reasonable. And in fact, our goal is valid inference on the parameter β^* regardless of (the quality of) $\hat{f}(\cdot)$.

3. An empirical investigation of the distribution of the test statistic

We consider a simple simulation setting, inspired by the *"Simulated Data: Continuous case"* section of Wang et al. (2020). They generate three datasets: a *training dataset* consisting of realizations of (Z, X, Y) used to train a machine learning model $\hat{f}(\cdot)$, a *labeled dataset* consisting of realizations of (Z, X, Y), and an *unlabeled dataset* consisting only of realizations of (Z, X); both the labeled and unlabeled datasets are used for inference[3]. They consider predictors $Z \in \mathbb{R}^4$ and response $Y \in \mathbb{R}$, and define the covariate $X \equiv Z_1$. In Wang et al. (2020)'s paper, the training, labeled, and unlabeled datasets each consist of 300 observations. Throughout this section, we keep the training sample size fixed at 300 observations, but vary the size of the labeled and unlabeled datasets.

As in Wang et al. (2020), we generate the training, labeled, and unlabeled datasets from the same partially linear additive model $Y = \tilde{\beta}_0 + \tilde{\beta}_1 Z_1 + \sum_{j=2}^{4} \tilde{\beta}_j g_j(Z_j) + \epsilon$. Their goal is to conduct inference on the marginal association between $X = Z_1$ and Y in a linear model. That is, their parameter of interest is

$$\beta_1^* = \arg\min_{\beta_1} \min_{\beta_0} \mathrm{E}[(Y - \beta_0 - \beta_1 Z_1)^2]. \tag{11}$$

---

3. Wang et al. (2020) refer to the labeled data set as "test" data, and to the unlabeled data as "validation" data.

Because the features are independent, we have that

$$\beta_1^* = \arg\min_{\beta_1} \min_{\beta_0,\beta_2,\beta_3,\beta_4} \mathrm{E}[(Y - \beta_0 - \beta_1 Z_1 - \sum_{j=2}^{4} \beta_j g_j(Z_j))^2] = \tilde{\beta}_1. \tag{12}$$

Thus, β_1^* is the marginal regression coefficient of Y onto $X = Z_1$, as well as the coefficient associated with $X = Z_1$ in the partially linear additive model used to generate the data. We consider two settings: one under the null ($\beta_1^* = 0$) and one under the alternative ($\beta_1^* = 1$).

We generate 3 training sets and fit a GAM to each training set, to obtain three fitted models $\hat{f}_1, \hat{f}_2, \hat{f}_3$. In each replicate of the simulation study, we generate a new labeled and unlabeled dataset as described above. Note that this differs from the simulation in Wang et al. (2020), which generates a new training set (and thus a different $\hat{f}$) in each replicate of the simulation study. We do this to focus on the properties of estimation and inference under a fixed $\hat{f}$, e.g. how AlphaFold would be used in practice. We perform a total of 1,000 simulation replicates.

To conduct prediction-based inference on β_1^*, both Wang et al. (2020) and Angelopoulos et al. (2023) rely on the claim that $(\hat{\beta}_1 - \beta_1^*)/\widehat{\mathrm{SE}}(\hat{\beta}_1) \xrightarrow{d} N(0,1)$. We consider the following versions of Wang et al. (2020):

1. *Proposal of Wang et al. (2020), with an analytical correction.* Apply Box 3 with the "analytical correction" (2) using a linear model for the regression model and a linear model for the relationship model.

2. *Proposal of Wang et al. (2020), with a "parametric bootstrap" correction.* Apply Box 3 with the "parametric bootstrap" correction presented in Algorithm 1 using a linear model for the regression model and a GAM for the relationship model.

3. *Proposal of Wang et al. (2020), with a "non-parametric bootstrap" correction.* Apply Box 3 with the "non-parametric bootstrap" correction presented in Algorithm 1 using a linear model for the regression model and a GAM for the relationship model.

We additionally consider the proposal of Angelopoulos et al. (2023):

4. *Proposal of Angelopoulos et al. (2023).* Apply Box 4 using a linear model.

Finally, we consider the following two approaches.

5. *Classical approach using only the labeled data.* Fit a linear model to $(\mathbf{y}_{\mathrm{lab}}, \mathbf{x}_{\mathrm{lab}})$.

6. *Naive approach.* Apply Box 1 using a linear model.

In Figure 1 we show the empirical distribution of $(\hat{\beta}_1 - \beta_1^*)/\widehat{\mathrm{SE}}(\hat{\beta}_1)$ under $H_0 : \beta_1^* = 0$, for increasing sample sizes. In the first three panels, we can see that *the asymptotic distribution of this test statistic for an arbitrary $\hat{f}$ does not follow a $N(0,1)$ for the naive and Wang et al. (2020) methods.* This is in line with our findings in Section 2. This is also true under the alternative $H_0 : \beta_1^* = 1$, as shown in Figure 2. On the other hand, for the method of Angelopoulos et al. (2023), this statistic converges in distribution to a $N(0,1)$ regardless of the choice of $\hat{f}$.
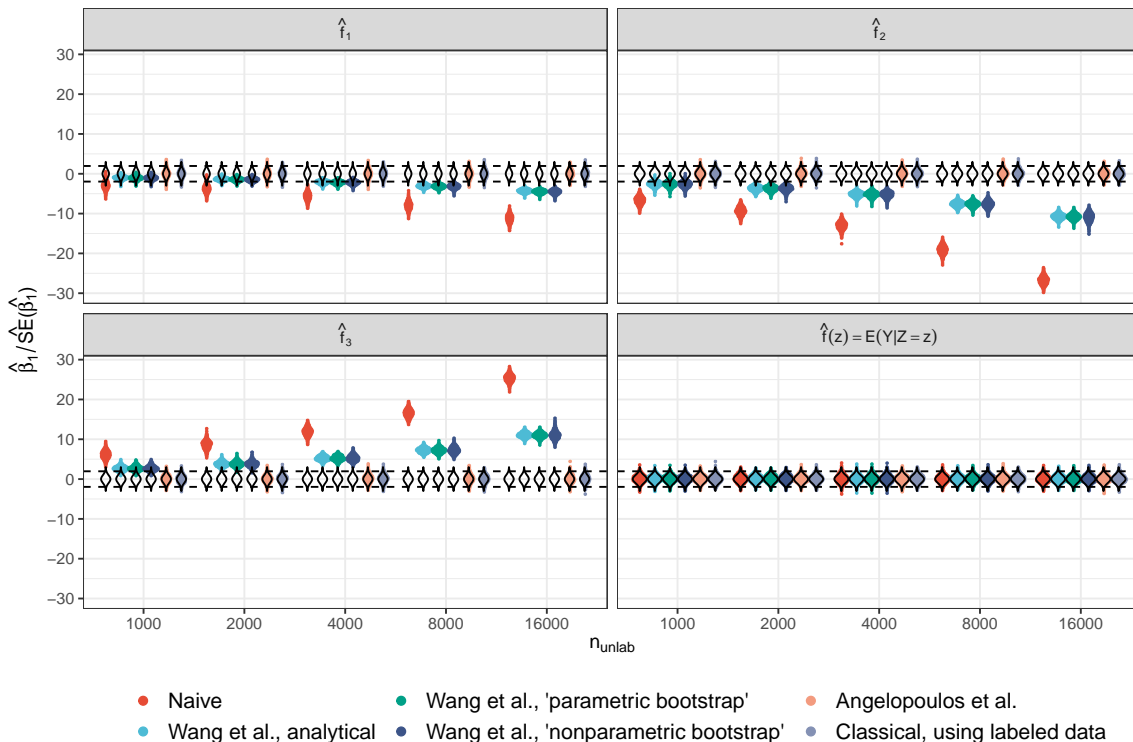
Figure 1: An examination of the distribution of $\hat{\beta}_1/\widehat{\mathrm{SE}}(\hat{\beta}_1)$ under $H_0 : \beta_1^* = 0$. For each of four different prediction models $\hat{f}(\cdot)$ (three trained GAMs and one true regression function), we display the empirical distribution of $\hat{\beta}_1/\widehat{\mathrm{SE}}(\hat{\beta}_1)$ as the sample sizes increase, with $n_{\mathrm{lab}} = 0.1 n_{\mathrm{unlab}}$. The $N(0,1)$ distribution is shown in black. The dashed black lines show the 0.025 and 0.975 quantiles of this distribution. The distributions of Wang et al. (2020)'s test statistics increasingly diverge from the $N(0,1)$ distribution as the sample sizes increase. The methods and simulation setup are described in Section 3.

In the last panel of Figures 1 and 2, we consider the extreme setting considered in Section 2.2, in which $\hat{f}(z) = \mathrm{E}[Y|Z = z]$. Here, the empirical distribution of the test statistic is approximately $N(0,1)$ for all methods.

We next performed testing and constructed confidence intervals under the assumption made by Wang et al. (2020) and Angelopoulos et al. (2023) that the test statistic asymptotically follows $N(0,1)$. We examined how the violation of this distributional assumption impacts type 1 error control and coverage in Figures 5 and 6 in the Appendix. We see that *the naive and Wang et al. (2020) methods do not control the type 1 error rate or have nominal coverage for an arbitrary $\hat{f}$,* and they become increasingly anti-conservative as the sample sizes increase. This can be explained by the increasing discrepancy between the assumed and true distributions of $(\hat{\beta}_1 - \beta_1^*)/\widehat{\mathrm{SE}}(\hat{\beta}_1)$ as the sample sizes increase, as previously seen in Figures 1 and 2. In fact, we can directly read off the type 1 error rate at level 0.05
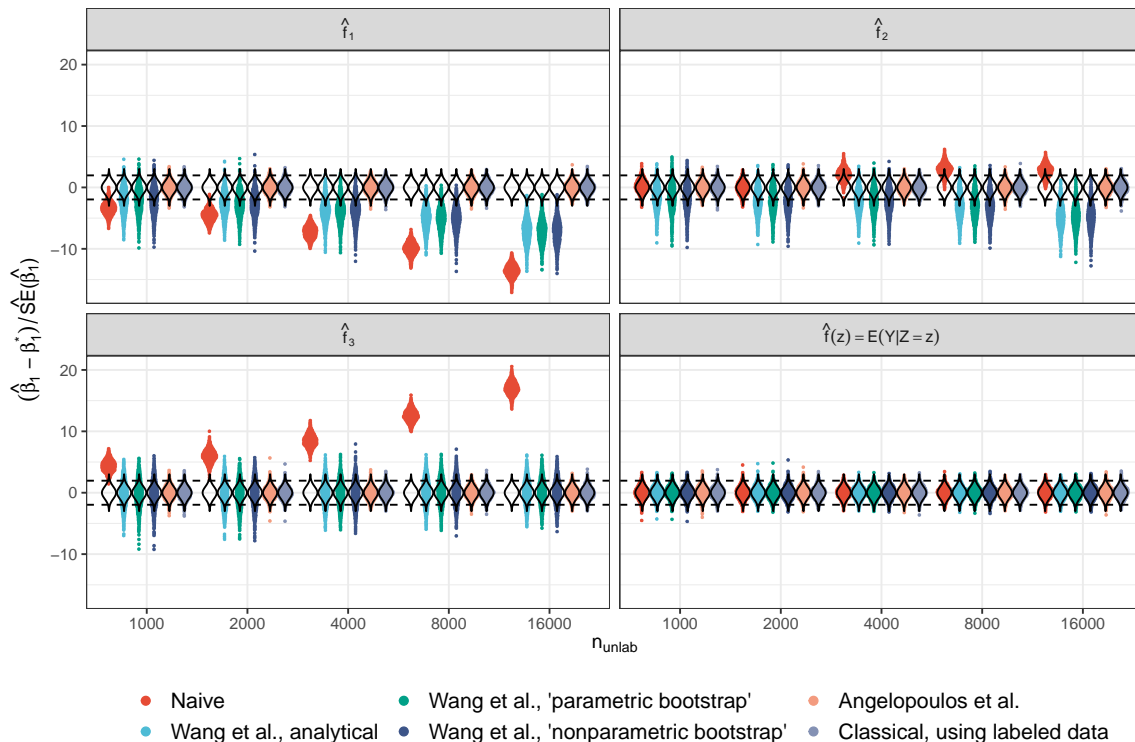
Figure 2: An examination of the distribution of $(\hat{\beta}_1 - \beta_1^*)/\widehat{\mathrm{SE}}(\hat{\beta}_1)$ when $\beta_1^* = 1$. For each of four different prediction models $\hat{f}(\cdot)$ (three trained GAMs and one true regression function), we display the empirical distribution of $(\hat{\beta}_1 - \beta_1^*)/\widehat{\mathrm{SE}}(\hat{\beta}_1)$ as the sample sizes increase, with $n_{\mathrm{lab}} = 0.1 n_{\mathrm{unlab}}$. The $N(0,1)$ distribution is shown in black. The dashed black lines show the 0.025 and 0.975 quantiles of this distribution. The distributions of Wang et al. (2020)'s test statistics increasingly diverge from the $N(0,1)$ distribution as the sample sizes increase. The methods and simulation setup are described in Section 3.

as the proportion of points falling outside the dashed lines in Figure 1. Similarly, coverage can be read as the proportion of points inside the dashed lines in Figure 2.

Because the true distribution of the test statistic matches the assumed one for the method of Angelopoulos et al. (2023) for any $\hat{f}$, this method gives correct type 1 error control and coverage in general. For the same reason, under the extreme setting, all methods have correct type 1 error control and coverage.

4. A direct replication of the simulation study of Wang et al. (2020)

In the previous section, we considered a simulation setting that was very similar to that in Wang et al. (2020), but differed in that we considered the same prediction models $\hat{f}$ across all simulation replicates. In this section, we instead directly replicate their simulation setting ("Simulated data; continuous case"), by generating a new training dataset in each

simulation replicate (resulting in a different $\hat{f}$ in each replicate). They considered a sample size of 300 for the training, labeled, and unlabeled datasets. In addition to replicating these results, we explore increasing the labeled and unlabeled dataset sizes.
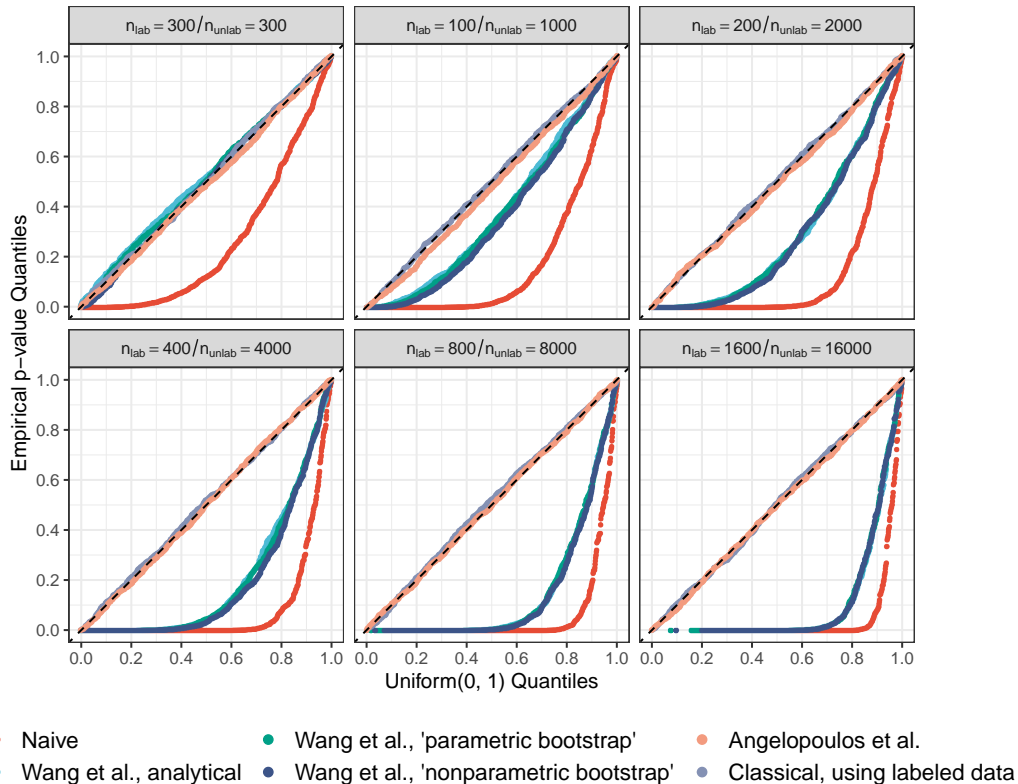

Figure 3: For data generated under $H_0 : \beta_1^* = 0$, quantile-quantile plots of the p-values across simulation replicates are displayed. The methods are described in Section 3 and the simulation setup is described in Section 4. Each panel corresponds to a different sample sizes of the labeled and unlabeled datasets used for inference. The bootstrap and analytical corrections considered by Wang et al. (2020) become increasingly anticonservative as the sample sizes increase. The classical approach, and that of Angelopoulos et al. (2023), are well-calibrated.

We first examine the type 1 error rate under the null hypothesis $H_0 : \beta_1^* = 0$ using each of the approaches described in Section 3. Quantile-quantile plots of the resulting $1,000$ p-values are shown in Figure 3. We see that in agreement with the findings in Wang et al. (2020), the naive approach does not control the type 1 error rate regardless of sample size. While it appears that when $n_{\text{lab}} = n_{\text{unlab}} = 300$ (first panel of Figure 3) the methods of Wang et al. (2020) have uniform p-values under the null (as also reported in Wang et al. (2020)), with other sample sizes this no longer holds and *the methods of Wang et al. (2020) fail to control the type 1 error rate.* As expected based on the previous section, Angelopoulos et al. (2023) controls type 1 error. This finding is in agreement with the theoretical results

in Angelopoulos et al. (2023), which hold for an arbitrary prediction function $\hat{f}(\cdot)$. Also as expected, the classical method controls type 1 error.
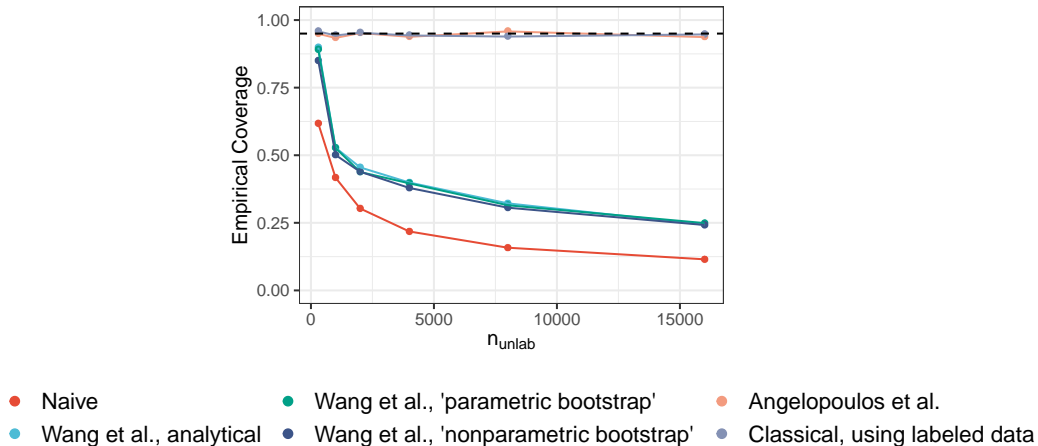

Figure 4: For data generated with $\beta_1^* = 1$, empirical coverage of 95% confidence intervals for each method across each simulation replicate, as the labeled and unlabeled sample sizes increase, with $n_{\text{lab}} = 0.1 n_{\text{unlab}}$. The methods are described in Section 3 and the simulation setup is described in Section 4.

Next we examine coverage of 95% confidence intervals under $\beta_1^* = 1$. We see from Figure 4 that the naive approach has coverage well below the nominal level. Again, *the "corrected" proposals of Wang et al. (2020) also fail to achieve the nominal coverage.* The problem becomes increasingly pronounced as the sample sizes of the data used for inference increase. By contrast, the classical method and the proposal of Angelopoulos et al. (2023) do achieve the nominal coverage, supporting the theory of Angelopoulos et al. (2023).

As explained in Section 2, the lack of inferential guarantees for the naive approach and the approach of Wang et al. (2020) is a direct consequence of the fact that $\hat{\beta}_1 \overset{p}{\nrightarrow} \beta_1^*$.

5. Discussion

In this paper, we found that the method of Angelopoulos et al. (2023) provides valid type 1 error control and coverage. By contrast, the methods proposed by Wang et al. (2020) do not provide appropriate inferential guarantees in the absence of very strong assumptions: for instance, under the extreme (and unrealistic) scenario where the prediction model is the true regression function. *Under this additional assumption, the naive approach also provides valid inference.* Furthermore, we see in our simulation study that simply assuming that the prediction model $\hat{f}(\cdot)$ was trained on data from the same population as the labeled and unlabeled data — an assumption made by Wang et al. (2020) — is not sufficient to achieve valid inference using their proposed methods.

Throughout, for simplicity we have assumed that we are conducting inference (i.e. Step 2 of Box 1) using a linear model. However, our conclusions — that the naive method and methods of Wang et al. (2020) result in invalid inference, because they target the incorrect

parameter — apply much more generally. The theory in Angelopoulos et al. (2023) shows that their approach applies to a wide variety of settings beyond linear regression, and has valid inferential properties.

Code Availability

Scripts to reproduce the results in this manuscript are available at `https://github.com/keshav-motwani/PredictionBasedInference/`. Our code is based on the code from Wang et al. (2020); we thank the authors for making it publicly accessible.

Acknowledgements

References

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000. URL `https://www.science.org/doi/abs/10.1126/science.adi6000`.

Samuel J Clark, Tyler McCormick, Zehang Li, and Jon Wakefield. InSilicoVA: A method to automate cause of death assignment for verbal autopsy. *arXiv preprint arXiv:1504.02129*, 2015.

Shannon E Ellis, Leonardo Collado-Torres, Andrew Jaffe, and Jeffrey T Leek. Improving the value of public RNA-seq expression data by phenotype prediction. *Nucleic Acids Research*, 46(9):e54–e54, 2018.

Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015.

Alexander Gusev, Kate Lawrenson, Xianzhi Lin, Paulo C Lyra Jr, Siddhartha Kar, Kevin C Vavra, Felipe Segato, Marcos AS Fonseca, Janet M Lee, Tanya Pejovic, et al. A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. *Nature Genetics*, 51(5):815–823, 2019.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

SA Khoury, D Massad, and T Fardous. Mortality and causes of death in Jordan 1995-96: assessment by verbal autopsy. *Bulletin of the World Health Organization*, 77(8):641, 1999.

Siruo Wang, Tyler H McCormick, and Jeffrey T Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020.

Appendix A. Necessity of consistency for β^*

We formally state the necessity of targeting the correct parameter in order to use the test statistic $(\hat{\beta}_j - \beta_j^*)/\widehat{\mathrm{SE}}(\hat{\beta}_j)$ for inference.

Lemma 1 *Suppose* $\widehat{\mathrm{SE}}(\hat{\beta}_j) = o_p(1)$ *and* $\hat{\beta}_j \overset{p}{\not\to} \beta_j^*$. *Then* $(\hat{\beta}_j - \beta_j^*)/\widehat{\mathrm{SE}}(\hat{\beta}_j)$ *does not converge in distribution.*

Proof Suppose, for the sake of contradiction, that $(\hat{\beta}_j - \beta_j^*)/\widehat{\mathrm{SE}}(\hat{\beta}_j)$ converges in distribution. Then $(\hat{\beta}_j - \beta_j^*)/\widehat{\mathrm{SE}}(\hat{\beta}_j) = O_p(1)$. Thus

$$\hat{\beta}_j - \beta_j^* = \frac{\hat{\beta}_j - \beta_j^*}{\widehat{\mathrm{SE}}(\hat{\beta}_j)}\widehat{\mathrm{SE}}(\hat{\beta}_j) = O_p(1)o_p(1) = o_p(1),$$

so $\hat{\beta}_j \overset{p}{\to} \beta_j^*$, which is a contradiction. ∎

Appendix B. Lack of consistency of analytical method of Wang et al. (2020)

In Sections 2.1 and 2.2, we analyzed the analytical correction as implemented in the code by Wang et al. (2020). We now analyze the analytical correction as described in the publication, which is also not consistent for the parameter of interest in general. In the publication, the estimate obtained from the analytical correction is defined as

$$\hat{\beta}_{\mathrm{Wang}}^{\mathrm{analytical}} = \hat{\gamma}_0(\mathbf{x}_{\mathrm{unlab}}^\top \mathbf{x}_{\mathrm{unlab}})^{-1}\mathbf{x}_{\mathrm{unlab}}^\top \mathbf{1}_{n_{\mathrm{unlab}}} + \hat{\gamma}_1 \hat{\beta}_{\mathrm{naive}}$$

where

$$\hat{\gamma}_1 = \frac{\widehat{\mathrm{Cov}}(\mathbf{y}_{\mathrm{lab}}, \hat{f}(\mathbf{z}_{\mathrm{lab}}))}{\widehat{\mathrm{Var}}(\hat{f}(\mathbf{z}_{\mathrm{lab}}))}$$

and

$$\hat{\gamma}_0 = \hat{\mathrm{E}}[\mathbf{y}_{\mathrm{lab}}] - \hat{\gamma}_1 \hat{\mathrm{E}}[\hat{f}(\mathbf{z}_{\mathrm{lab}})].$$

Thus as n_{lab} increases,

$$\hat{\gamma}_1 \overset{p}{\to} \frac{\mathrm{Cov}(Y, \hat{f}(Z))}{\mathrm{Var}(\hat{f}(Z))} \equiv \gamma_1$$

and

$$\hat{\gamma}_0 \overset{p}{\to} \mathrm{E}[Y] - \gamma_1 \mathrm{E}[\hat{f}(Z)] \equiv \gamma_0.$$

Thus as n_{lab} and n_{unlab} increase,

$$\hat{\beta}_{\mathrm{Wang}}^{\mathrm{analytical}} \overset{p}{\to} \gamma_0 \mathrm{E}[XX^\top]^{-1}\mathrm{E}[X] + \gamma_1 \mathrm{E}[XX^\top]^{-1}\mathrm{E}[X\hat{f}(Z)],$$

which is not equal to β^* in general.

In the extreme setting (described in Section 2.2), in which $\hat{f}(z) = \mathrm{E}[Y|Z = z]$, we have that

$$\gamma_1 = \frac{\mathrm{Cov}(Y, \hat{f}(Z))}{\mathrm{Var}(\hat{f}(Z))} = \frac{\mathrm{Cov}(Y, \mathrm{E}[Y|Z])}{\mathrm{Var}(\mathrm{E}[Y|Z])} = \frac{\mathrm{Var}(\mathrm{E}[Y|Z])}{\mathrm{Var}(\mathrm{E}[Y|Z])} = 1$$

and

$$\gamma_0 = \mathrm{E}[Y] - \gamma_1 \mathrm{E}[\hat{f}(Z)] = \mathrm{E}[Y] - \mathrm{E}[\mathrm{E}[Y|Z]] = 0.$$

Thus using the same argument as for the naive estimator, this analytical correction is consistent for β^* in this extreme setting.

Appendix C. Inferential consequences of wrong distribution

In this section, we examine how violation of the distributional assumption impacts type 1 error control and coverage in the simulations in Section 3. In Figure 5, we see that the methods proposed by Wang et al. (2020) do not control the type 1 error rate for arbitrary $\hat{f}(\cdot)$; the situation gets worse as the sample size increases. However, the method proposed by Angelopoulos et al. (2023) does control the type 1 error rate. Wang et al. (2020) controls the type 1 error rate if the machine learning model is the true regression function $\hat{f}(z) = \mathrm{E}[Y|Z = z]$; of course, such a perfect machine learning model is unattainable in practice.

In Figure 6, we see that the methods proposed by Wang et al. (2020) do not attain the nominal coverage for an arbitrary $\hat{f}(\cdot)$, whereas the proposal of Angelopoulos et al. (2023) does attain the nominal coverage. The naive method and the proposals of Wang et al. (2020) have appropriate coverage when $\hat{f}(z) = \mathrm{E}[Y|Z = z]$; again, this is unrealistic in practice.
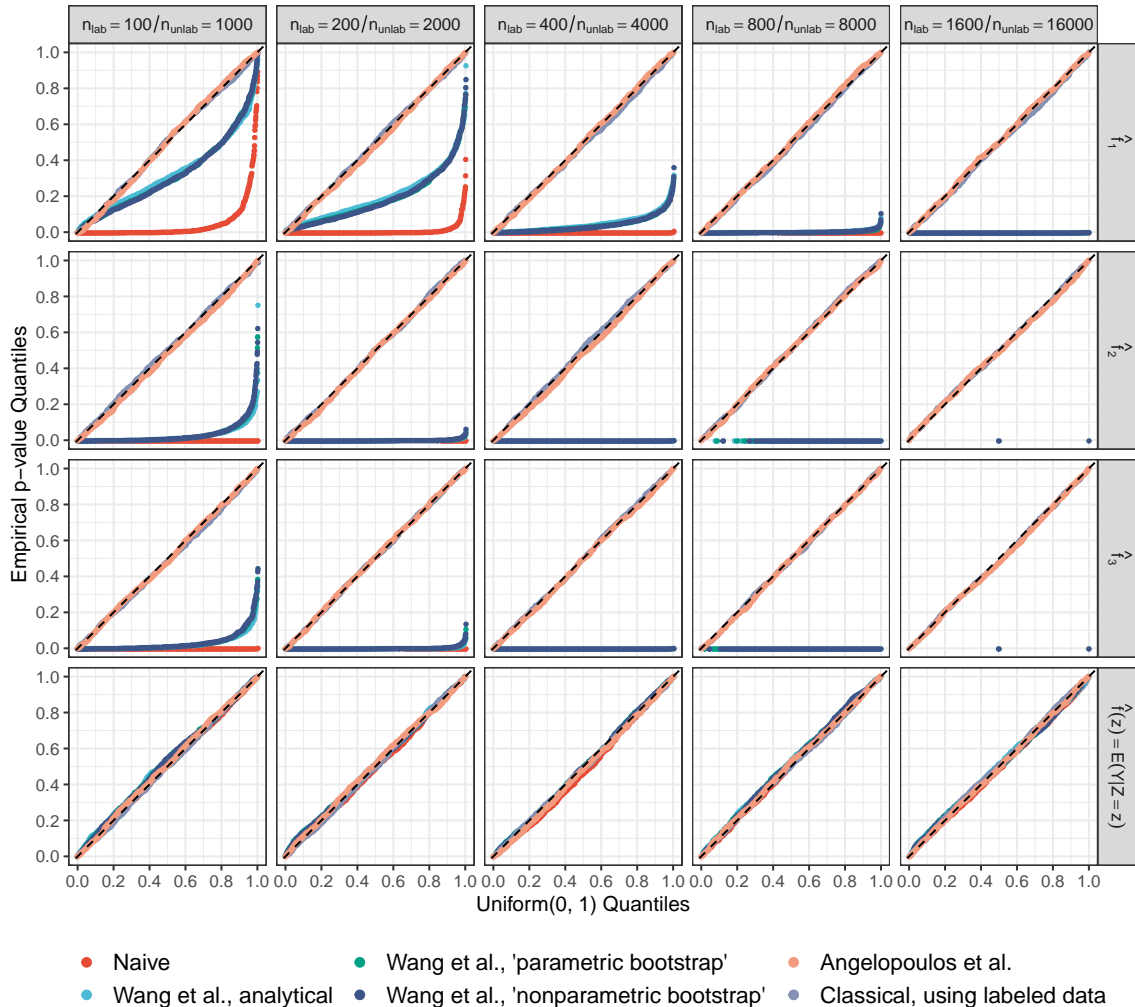
Figure 5: For labeled and unlabeled datasets generated under $H_0 : \beta_1^* = 0$, quantile-quantile plots of the p-values across replicates of the modified simulation study are displayed for each of the four $\hat{f}(\cdot)$'s considered. The methods and simulation setup are described in Section 3. Each panel corresponds to different sample sizes of the labeled and unlabeled datasets used for inference. The naive method and the bootstrap and analytical corrections considered by Wang et al. (2020) become increasingly anticonservative as the sample sizes increases, unless the machine learning model is perfect, i.e. $\hat{f}(z) = \mathrm{E}[Y|Z = z]$.
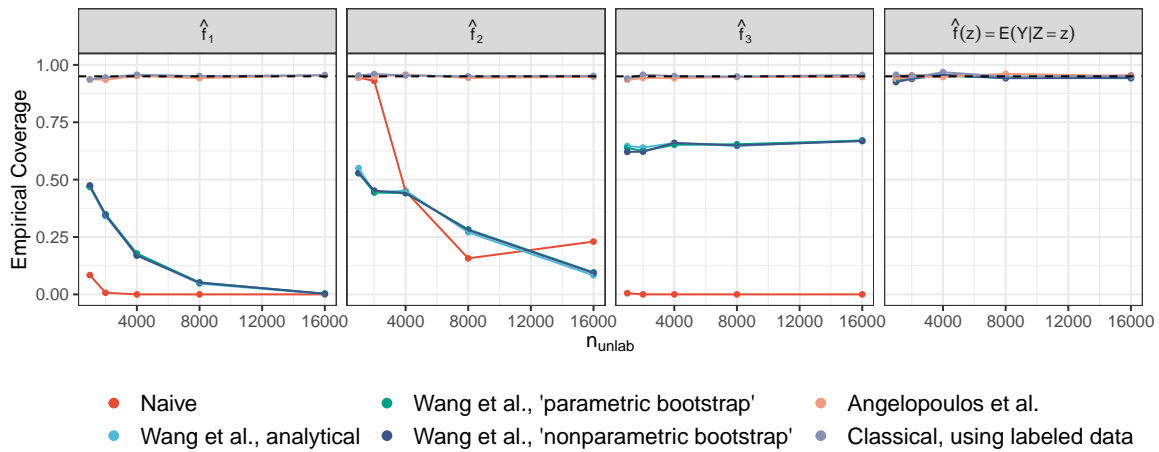
Figure 6: For labeled and unlabeled datasets generated with $\beta_1^* = 1$, empirical coverage of 95% confidence intervals for each method across each simulation replicate, for each of the four $\hat{f}(\cdot)$'s considered, as the labeled and unlabeled sample sizes increase, with $n_{\text{lab}} = 0.1 n_{\text{unlab}}$. The methods and simulation setup are described in Section 3.